A Survey on Knowledge Distillation of Large Language Models

Year: 2024 | *Citations: 218* | *Authors: Xiaohan Xu, Ming Li, Chongyang Tao*

Abstract

In the era of Large Language Models (LLMs), Knowledge Distillation (KD) emerges as a pivotal methodology for transferring advanced capabilities from leading proprietary LLMs, such as GPT-4, to their open-source counterparts like LLaMA and Mistral. Additionally, as open-source LLMs flourish, KD plays a crucial role in both compressing these models, and facilitating their self-improvement by employing themselves as teachers. This paper presents a comprehensive survey of KD's role within the realm of LLM, highlighting its critical function in imparting advanced knowledge to smaller models and its utility in model compression and self-improvement. Our survey is meticulously structured around three foundational pillars: \textit{algorithm}, \textit{skill}, and \textit{verticalization} -- providing a comprehensive examination of KD mechanisms, the enhancement of specific cognitive abilities, and their practical implications across diverse fields. Crucially, the survey navigates the intricate interplay between data augmentation (DA) and KD, illustrating how DA emerges as a powerful paradigm within the KD framework to bolster LLMs' performance. By leveraging DA to generate context-rich, skill-specific training data, KD transcends traditional boundaries, enabling open-source models to approximate the contextual adeptness, ethical alignment, and deep semantic insights characteristic of their proprietary counterparts. This work aims to provide an insightful guide for researchers and practitioners, offering a detailed overview of current methodologies in KD and proposing future research directions. Importantly, we firmly advocate for compliance with the legal terms that regulate the use of LLMs, ensuring ethical and lawful application of KD of LLMs. An associated Github repository is available at https://github.com/Tebmer/Awesome-Knowledge-Distillation-of-LLMs.